

Multichannel Identification and Nonnegative Equalization for Dereverberation and Noise Reduction based on Convolutional Transfer Function

Xiaofei Li, Sharon Gannot, Laurent Girin and Radu Horaud

Abstract—This paper addresses the problems of blind multichannel identification and equalization for joint speech dereverberation and noise reduction. The time-domain cross-relation method is hardly applicable for blind room impulse response identification due to the near-common zeros of the long impulse responses. We extend the cross-relation method to the short-time Fourier transform (STFT) domain, in which the time-domain impulse response is approximately represented by the convolutional transfer function (CTF) with much less coefficients. For the oversampled STFT, CTFs suffer from the common zeros caused by the non-flat frequency response of the STFT window. To overcome this, we propose to identify CTFs using the STFT framework with oversampled signals and critically sampled CTFs, which is a good trade-off between the frequency aliasing of the signals and the common zeros problem of CTFs. The identified complex-valued CTFs are not accurate enough for multichannel equalization due to the frequency aliasing of the CTFs. Thence, we only use the CTF magnitudes, which leads to a nonnegative multichannel equalization method based on a nonnegative convolution model between the STFT magnitude of the source signal and the CTF magnitude. Compared with the complex-valued convolution model, this nonnegative convolution model is shown to be more robust against the CTF perturbations. To recover the STFT magnitude of the source signal and to reduce the additive noise, the ℓ_2 -norm fitting error between the STFT magnitude of the microphone signals and the nonnegative convolution is constrained to be less than a noise power related tolerance. Meanwhile, the ℓ_1 -norm of the STFT magnitude of the source signal is minimized to impose the sparsity.

I. INTRODUCTION

This work addresses the problem of joint blind multichannel dereverberation and noise reduction. The goal is to remove the reverberation and noise from the microphone signals to improve the speech intelligibility for both human listening and machine recognition. The output of a dereverberation system can include some early reflections, since early reflections do not deteriorate the speech quality and speech intelligibility [1].

Multichannel dereverberation can be processed by different techniques. Spectral enhancement techniques remove the late reverberation by spectral subtraction. Many techniques have

been proposed to estimate the power spectral density (PSD) of late reverberation, such as convolutional transfer function (CTF) based statistical model [2], maximum likelihood [3], [4], coherent-to-diffuse power ratio (CDR) [5], where [3], [4], [5] also take the additive noise into account. The weighted prediction error (WPE) method [6], [7], [8] first estimates the late reverberation by filtering the microphone signals with linear prediction filters, and then subtract it from the microphone signals. Noise suppression is further integrated in [9], [10], [11]. Probabilistic techniques use expectation-maximization (EM) algorithm to maximize the likelihood of a generative model of the noisy microphone signals, such as [12], [13] using relative early transfer function, and [14], [15] using CTF. Multichannel equalization techniques first blindly identify the channel filters [16], then apply the inverse filtering on the microphone signals [17], [18], [19], [20]. Note that the above four classes of methods are broadly summarized, and only a few references are named. Some of them have some common characteristics, for example, a probabilistic model is also used in some spectral enhancement techniques [3], [4] and in the WPE methods.

The focus of this paper is multichannel equalization technique. For a single-input multiple-output (SIMO) system, the blind channel identification can be carried out based on the second-order statistics, such as the subspace method [21] or the cross-relation method [16]. The cross-relation method identifies the channel filters by detecting the eigenvector corresponding to the unique zero eigenvalue of the covariance matrix of microphone signals. A noise subspace method proposed in [22] exploited the multiple eigenvectors corresponding to the zero eigenvalues of the over-modeled covariance matrix. These noise-subspace based methods, especially the cross-relation method, are vulnerable to additive noise and the filter length determination error. Thence they require a prior knowledge of the exact filter length. However, in acoustic dereverberation, the room impulse response (RIR) is a time sequence with variance exponentially decaying to zero. The filter length is hardly measureable, namely the truncation point is difficult to determine. For the case of small tails, [23] proposed a channel under-modeling method that only considers the significant part of the filters, and a rank detection method was proposed in [24] to determine the length of the significant part. However, these methods are applicable only when a noticeable gap exists between the significant part and the small tails, which is obviously not the case of

Xiaofei Li and Radu Horaud are with INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France. E-mail: first.last@inria.fr

Sharon Gannot is with Bar Ilan University, Faculty of Engineering, Israel. E-mail: Sharon.Gannot@biu.ac.il

Laurent Girin is with INRIA Grenoble Rhône-Alpes and with Univ. Grenoble Alpes, GIPSA-lab, Grenoble, France. E-mail: laurent.girin@gipsa-lab.grenoble-inp.fr

This work was supported by the ERC Advanced Grant VHIA #340113.

RIRs. Based on the least mean squares method, the frequency-domain adaptive cross-relation method was proposed in [25], and was applied to speech separation and dereverberation in [26]. One of the identifiability conditions of the second-order statistics-based methods is that the multiple channels are co-prime, namely they do not share any common zeros. It is shown in [27] that a large number of near-common zeros exists for long channel filters, which deteriorates the performance of channel identification. A forced spectral diversity algorithm [28], [29] was proposed to mitigate the near-common zeros problem. A ℓ_1 -norm sparse learning was exploited in [30], [31] by assuming that RIR is sparse, which is valid for the early reflections, whereas it is less valid for the late reverberation.

For multichannel equalization (or inverse filtering) using the known channel filters, ideally, the classical multiple-input/output inverse theorem (MINT) method [17] can perfectly recover the source signal. In MINT, the inverse filter is obtained by equalizing the channel filters, targeting an impulse function, and is applied to the microphone signals. However, MINT is sensitive to filter perturbations and to additive noise in the microphone signals. To improve the robustness of MINT to RIR perturbations, many techniques have been proposed, preserving not only the direct-path impulse response but also the early reflections, such as channel shortening [32], infinity- and p -norm optimization-based channel shortening/reshaping [33], partial MINT [34], and relaxed multichannel least squares [19]. The energy of the inverse filter was used in [18] as a regularization term to avoid the amplification of filter perturbations and microphone noise. For joint dereverberation and noise reduction, the output noise power was used in [20] as a regularization term. Without explicitly estimating the inverse filter, a wide-band Lasso method was proposed in [35] for both source separation and dereverberation. The source signals are estimated by minimizing a time-domain ℓ_2 -norm fitting cost between the microphone signals and the mixture model involving the unknown source signals. Importantly, the ℓ_1 -norm of the short-time Fourier transform (STFT) domain source signals is added to the mixture fitting cost as a regularization term to impose the sparsity of speech in the time-frequency domain. This regularization term was then adapted to MINT in [36]. In the presence of additive noise, the ℓ_1 -norm regularization is able to reduce the noise in the recovered source signals. However, the regularization factor is difficult to set even if the noise power is known. To overcome this, a more flexible scheme is proposed in [37] that relaxes the ℓ_2 -norm mixture fitting cost to the noise level.

The channel identification and equalization methods mentioned above are all performed in the time domain. In the present paper, we consider dereverberation in the STFT domain. To represent the time-domain convolution in the STFT domain, especially for the long filter case, cross-band filters were introduced in [38]. To simplify the analysis, the CTF approximation can be adopted, e.g. in [39], only using the band-to-band convolution and ignoring the cross-band filters. The convolution between the time-domain source signal and RIR is approximated by the convolution between the source signal STFT coefficients and the CTF. The advantages of

the CTF over time-domain representation are i) CTFs are much shorter than RIRs, consequently are likely to have less near-common zeros, ii) the sparsity of speech spectra can be more easily exploited directly in the STFT domain. An EM algorithm based on the CTF convolution was proposed in [14], [15], [40] to iteratively estimate CTF and the source signal. In [41], [42], [43], a nonnegative approximation is assumed and demonstrated, namely the STFT magnitude of the source image is approximated by the convolution between the STFT magnitude of source signal and the CTF magnitude. Based on this nonnegative model, the tensor factorization in [41] and the iterative auxiliary functions in [42] were used for dereverberation, and the iterative multiplicative update was used in [43] for joint dereverberation and denoising. In parallel to the CTF model developments, it is well known that the STFT can be interpreted as a filter-bank decomposition [44]. Adaptive filtering in subbands has been widely studied [45], [46] and applied to acoustic echo cancellation, while the subband blind channel identification has been rarely studied. In [22], the noise subspace method was applied in subbands, however it was not applied to real scenarios. For multichannel equalization, several variants of subband MINT were proposed based on filter banks [47], [48], [49], [50]. In our previous work [51], a CTF-Lasso method was proposed for source separation following the spirit of the wide-band Lasso [35].

This paper proposes a blind CTF identification method, and a nonnegative CTF-based multichannel equalization method for joint dereverberation and noise reduction. In each frequency band, the cross-relation method [16] is extended for CTF identification. First, the influence of the STFT configuration is analyzed. The frequency response of the short-time STFT window, e.g. Hamming window, is wider than the bandwidth of one frequency band, which means overlap exists among adjacent frequency bands. In addition, the main lobe of the frequency response is not flat, and the frequency region close to the margin of the main lobe have a magnitude close to zero. The CTF model disregards the cross-band filters, thus suffers from an under-modeling error, which depends on the subband overlap among adjacent frequencies. For the oversampling case, namely the STFT frame step is smaller than the STFT frame length, zeros exist in the frequency response of the CTF due to the non-flat frequency response of STFT window. These zeros are common to all channels and thus problematic for the cross-relation method. This can be avoided by critical sampling, which however leads to a severe frequency aliasing. Second, to achieve a good trade-off, the following scheme is used. The signal STFT coefficients are oversampled to avoid the frequency aliasing, and the CTFs are forced to be critically sampled to avoid the common zeros problem. Third, instead of eigendecomposition, similar to [30], we estimate the CTFs by solving a least-square problem that the cross-relation cost is minimized and the summation of the first tap of CTFs is constrained to equal one. This method is robust to the noise interference and to the filter length determination error. Using the proposed method, the identified complex-valued CTFs are not accurate enough for multichannel equalization due to the frequency aliasing of

the CTFs. Therefore, only the CTF magnitudes, and consequently a nonnegative convolution model, are used, which is shown to be less sensitive to the CTF perturbations than the complex-valued convolution model. This leads to the following nonnegative multichannel equalization method. In the same spirit of [37], an optimization problem is adopted for joint dereverberation and noise reduction. The STFT magnitude of source signal is recovered by minimizing its ℓ_1 -norm to impose the sparsity of the speech spectra. To reduce the additive noise, the ℓ_2 -norm fitting cost between the STFT magnitude of microphone signals and the nonnegative convolution model involving the unknown STFT magnitude of source signal is constrained to be less than a tolerance with respect to the noise power. The primal-dual interior-point method (PDIPM) [52] is used to solve this optimization problem. Finally, the phase of one of the microphone signals is combined with the estimated STFT magnitude of source signal, and the time-domain signal is obtained by inverse STFT. Overall, the main contributions of this work are the followings: i) we analyze the influence of the STFT configuration on signal reconstruction, CTF approximation and the common zeros issue, ii) in the oversampled STFT framework, we propose to force the channel filters to be critically sampled to avoid the common zeros problem, and iii) in the spirit of [37], a nonnegative multichannel equalization is proposed based on the nonnegative CTF convolution.¹

The remainder of this paper is organized as follows. The blind channel identification in the STFT domain is presented in Section II. The nonnegative multichannel equalization method for dereverberation and noise reduction is presented in Section III. Experiments with binaural simulation data and with multichannel real recordings are presented in Section IV and V, respectively. Section VI concludes the work.

II. STFT-DOMAIN CHANNEL IDENTIFICATION

We consider a two channel system. In the time domain, the noise-free microphone signals $x(n)$ and $y(n)$ are

$$x(n) = s(n) \star a(n), \quad y(n) = s(n) \star b(n), \quad (1)$$

where \star denotes convolution, $s(n)$ is a non-stationary source signal, e.g., speech, and $a(n)$ and $b(n)$ are the RIRs.

A. Problem Formulation in the STFT Domain

The STFT representation of the signal $x(n)$ is denoted as $x_{p,k}$, where p and k denote the frame and frequency indices, respectively. The cross-band filter model consists in representing the STFT coefficient $x_{p,k}$ as a summation over multiple convolutions (between the STFT-domain source signal $s_{p,k}$ and filter $a_{p',k'}$) across frequency bins. Mathematically, (1) can be written in the STFT domain as

$$x_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'=-C}^{Q-1} s_{p-p',k'} a_{p',k,k'}, \quad (2)$$

¹Supplemental materials are available online at: <https://team.inria.fr/perception/research/ctf-dereverberation>

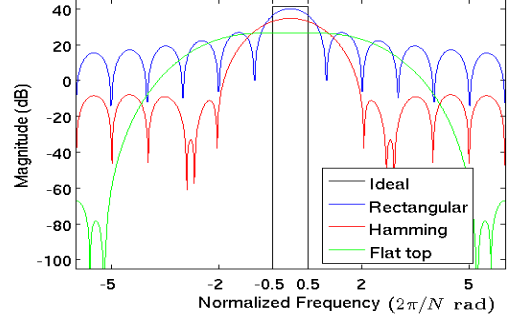


Fig. 1: The frequency response of STFT windows.

where N denotes the frame (window) length. Let L denote the frame step. If $L < N$, then $a_{p',k,k'}$ is non-causal, with $C = \lceil N/L \rceil - 1$ non-causal coefficients, where $\lceil \cdot \rceil$ denotes ceiling operator. The number of causal filter coefficients Q is related to the reverberation time. Let $\tilde{w}(n)$ and $w(n)$ denote the STFT analysis window and synthesis window, respectively. The STFT-domain impulse response $a_{p',k,k'}$ is related to the time-domain impulse response $a(n)$ by:

$$a_{p',k,k'} = (a(n) \star \zeta_{k,k'}(n))|_{n=p'L}, \quad (3)$$

which represents the convolution with respect to the time index n evaluated at frame steps, with

$$\zeta_{k,k'}(n) = e^{j\frac{2\pi}{N}k'n} \sum_{m=-\infty}^{+\infty} \tilde{w}(m) w(n+m) e^{-j\frac{2\pi}{N}m(k-k')}.$$

To simplify the analysis, we consider the CTF approximation, i.e., only band-to-band filter with $k = k'$ is considered

$$x_{p,k} \approx \sum_{p'=-C}^{Q-1} s_{p-p',k} a_{p',k} = s_{p,k} \star a_{p,k}. \quad (4)$$

Similarly, we have $y_{p,k} \approx s_{p,k} \star b_{p,k}$. To identify the filters $a_{p,k}$ and $b_{p,k}$, the cross-relation between the two channels

$$x_{p,k} \star b_{p,k} = s_{p,k} \star a_{p,k} \star b_{p,k} = y_{p,k} \star a_{p,k} \quad (5)$$

can be used. This relation was originally proposed for the time-domain filter identification and here is extended to the CTF domain. The conditions that this identification problem has a unique solution are given in [16], namely that i) the source signal $s_{p,k}$ should fully excite the filters, and that ii) the two filters $a_{p,k}$ and $b_{p,k}$ are co-prime, i.e. they do not share any common zeros. Otherwise, the common zeros are unidentifiable, since in the identified filters, a common zero can be replaced with any other zero without violating the cross-relation (5). In practice, the first condition can be satisfied by increasing the length of the speech signal and thus enriching the frequency content. The second condition is related to the STFT configuration. Prior to the detailed filter identification algorithm, below we analyze the influence of the STFT configuration on signal reconstruction, CTF approximation and the common zeros issue.

B. Analysis of STFT Configuration

Let $\check{W}(\omega)$ denote the frequency response of $\check{w}(n)$ obtained by applying the discrete-time Fourier transform with respect to n , where ω denotes the angular frequency. Fig. 1 shows the frequency response of three typical windows, i.e. rectangular, Hamming and flat-top windows, which have a main lobe width of about $2B$, $4B$ and $10B$, respectively, where $B = 2\pi/N$ is the bandwidth of one STFT frequency bin. In addition, the ideal low-pass filter with bandwidth B is shown with a black rectangle. Without loss of generality, the synthesis window $w(n)$ is assumed to be identical to the analysis window.

The filter-bank interpretation of STFT is that the time domain signal is first modulated (frequency shifted) by $e^{-j\frac{2\pi}{N}kn}$, then low-pass filtered with the analysis window $\check{w}(n)$, and downsampled by L . The downsampling operation folds the frequency content with the period of $2\pi/L$. For the ideal filter, there is no frequency aliasing up to the critical downsampling. However for the practical case, to have a low frequency aliasing, we should use a window with a high side-lobe attenuation, and choose a small L to make $2\pi/L$ not less than the width of main lobe. For example, the ideal, Hamming and flat-top windows have a high-side lobe attenuation, and L should be not larger than $L_{max} = N$, $N/4$ and $N/10$, respectively.

To illustrate the reliability of the CTF approximation, we analyze the significance of the cross-band filters for various windows. According to (3), we have the undecimated STFT-domain filter $\check{a}_{n,k,k'}$ by setting $L = 1$, which has the frequency response [38]:

$$\check{A}_{k,k'}(\omega) = A(\omega)\check{W}\left(\omega - \frac{2\pi}{N}k\right)W\left(\omega - \frac{2\pi}{N}k'\right) \quad (6)$$

where $A(\omega)$, $\check{W}(\omega)$ and $W(\omega)$ are the frequency responses of $a(n)$, $\check{w}(n)$ and $w(n)$, respectively. The product of $\check{W}(\omega - \frac{2\pi}{N}k)$ and $W(\omega - \frac{2\pi}{N}k')$ indicates the power of the cross-band filters. For the ideal filter, the cross-band filters $\check{A}_{k,k'}(\omega)$ with $k' \neq k$ are equal to zero, which means that the CTF is a perfect STFT-domain representation of the time-domain filter. However, for the three practical windows discussed above, it can be deduced from Fig. 1 that $\check{A}_{k,k'}(\omega)$ with $k' \neq k$ are not zero. The CTF approximation error is the power of the cross-band filters with $k' \neq k$ relative to the power of the band-to-band filter. Consider $\omega = 0$, it can be deduced from Fig. 1 that the rectangular window has the smallest approximation error among the three windows, since its overlap between $\check{W}(0)$ and $W(\frac{2\pi}{N}k')$ is the smallest. In contrast, the flat-top window has the largest approximation error.

The frequency response of the band-to-band filter $a_{p,k}$ with the decimation factor L is

$$\check{A}_{k,k}(\omega)_{\downarrow L} = \frac{1}{L} \sum_{l=0}^{L-1} \check{A}_{k,k}\left(\frac{\omega}{L} - \frac{2\pi}{L}l\right), \quad (7)$$

where $\check{A}_{k,k}(\omega)$ is defined in (6) with $k' = k$. Without loss of generality, we consider the case that N/L is an integer, and

the frequency band k is an integer multiple of N/L . Then (7) is simplified as

$$\check{A}_{k,k}(\omega)_{\downarrow L} = \frac{1}{L} A\left(\frac{\omega}{L} - \frac{2\pi}{N}k\right) \check{W}\left(\frac{\omega}{L}\right) W\left(\frac{\omega}{L}\right). \quad (8)$$

The filter $\check{W}(\frac{\omega}{L})$ involves only the main lobe if $L = L_{max}$, and involves some side lobes if $L < L_{max}$. For the ideal filter, the main lobe is flat, namely $\check{W}(\frac{\omega}{L})$ is flat when $L = L_{max} = N$. For the three practical windows, the magnitudes of $\check{W}(\frac{\omega}{L})$ and $W(\frac{\omega}{L})$ are close to zero in the side lobe, and in the marginal region of the main lobe. This close-to-zero-magnitude region is caused by the STFT window, and thence is also present in the frequency response of $b_{p,k}$. These common zeros of the two channels are problematic for the cross-relation method (5). The flat-top window has the least common zeros among the three windows.

To summarize, extending the time-domain cross-relation method to the STFT domain is not a trivial task. It suffers from the problems of frequency aliasing, CTF approximation and common zeros. The windows (low-pass filters) and the frame step (decimation factor) are crucial for circumventing these problems. Briefly, the frequency aliasing can be suppressed by using a window with a high side-lobe attenuation, and a small frame step. A small CTF approximation error requires the window to have a narrow main lobe. To avoid the common zeros, the window should have a flat main lobe. Otherwise, a larger frame step (even critical sampling) is needed to have the nearly flat frequency response of the CTF. Unfortunately, all these requirements can only be satisfied by the ideal filter with critical sampling.

In this work, to achieve a good trade-off between these requirements, we use the following STFT configuration. The Hamming window with frame step $L = N/4$ is adopted for the STFT of the signals, which has negligible frequency aliasing. Hamming window has a moderate CTF approximation error among the three windows. In addition, to avoid the common zeros problem, we propose to force the STFT-domain filters $a_{p,k}$ and $b_{p,k}$ to be critically sampled, i.e. we set $L_f = N$, where L_f denotes the frame step of the STFT-domain filters. The details will be presented in the next subsection. As for the window length, a large one leads to a small number of CTF taps which is beneficial to channel identification, but also brings a large number of early reflections to the recovered source signal as will be shown in the next subsection. Thence, a good trade-off should be made, e.g. 64 ms in this work.

C. Channel Identification

Since the channel identification algorithm is applied frequency-wise, hereafter the frequency index k is omitted unless necessary. In (4) and (5), the frame index p corresponds to the frame step $L = N/4$, with $p \in [0, P-1]$ for the signals s_p , x_p and y_p , and $p \in [-C, Q]$ for the filters a_p and b_p . Denote the filters in vector form as \mathbf{a} and \mathbf{b} . To avoid the common zeros problem, we further downsample the filters by a factor of 4. The downsampled filters $a_{p\downarrow 4}$ and $b_{p\downarrow 4}$ correspond to the critical sampling, thus have a larger

frequency aliasing than the original filters, and no longer have non-causal coefficients. The downsampled filters start with the tap 0, and have a length of $\tilde{Q} = \lceil Q/4 \rceil$. Let us write them in vector form as $\tilde{\mathbf{a}} = [a_0, a_4, \dots, a_{4(\tilde{Q}-1)}]^\top$, $\tilde{\mathbf{b}} = [b_0, b_4, \dots, b_{4(\tilde{Q}-1)}]^\top$, where $^\top$ is the transpose operator.

Define the convolution matrix \mathbf{X} from the signal x_p as

$$\mathbf{X} = \begin{bmatrix} x_0 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ x_p & x_{p-4} & \ddots & \ddots & x_{p-4(\tilde{Q}-1)} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ x_{P-1} & x_{P-5} & \cdots & \cdots & x_{P-1-4(\tilde{Q}-1)} \end{bmatrix} \quad (9)$$

with the size of $P \times \tilde{Q}$. The row number is the frame number of the oversampled signal, while the column number is the length of the critically sampled filter. Thence $\mathbf{X}\tilde{\mathbf{b}}$ is the convolution between the oversampled signal and the critically sampled filter interpolated with zeros by a factor of 4. Alternatively, we can say that 3/4 of the original oversampled CTF coefficients are forced to be zero. The convolution matrix \mathbf{Y} is defined from y_p following the same principle. Then (5) can be rewritten as $\mathbf{X}\tilde{\mathbf{b}} = \mathbf{Y}\tilde{\mathbf{a}}$, or $\mathbf{Z}\tilde{\mathbf{c}} = \mathbf{0}$, where $\mathbf{Z} = [\mathbf{Y}, -\mathbf{X}]$, $\tilde{\mathbf{c}} = [\tilde{\mathbf{a}}^\top, \tilde{\mathbf{b}}^\top]^\top$, and $\mathbf{0}$ is a vector with all entries equal to 0. In [16], the filter vector $\tilde{\mathbf{c}}$ is estimated by taking an eigenvector of \mathbf{Z} corresponding to a zero eigenvalue. This method is only reliable in the case of exactly known filter length. In addition, even if the filter length is known, the one-dimensional null space of \mathbf{Z} could be easily contaminated even by a mild noise interference.

Instead, we estimate the filter vector $\tilde{\mathbf{c}}$ by solving the following least-square problem:

$$\min \|\mathbf{Z}\tilde{\mathbf{c}}\|^2, \quad \text{s.t. } \mathbf{g}^\top \tilde{\mathbf{c}} = 1, \quad (10)$$

where $\|\cdot\|$ denotes ℓ_2 -norm, and \mathbf{g} is a constant vector

$$\mathbf{g} = [1, \underbrace{0, \dots, 0}_{\tilde{Q}-1}, 1, \underbrace{0, \dots, 0}_{\tilde{Q}-1}]^\top. \quad (11)$$

Here we constrain the sum of the first entries of the two filters to 1, i.e. $a_0 + b_0 = 1$. Constraining the scale of the first entries will remove the delay ambiguity, namely the direct-path responses are enforced to start with the first entries. The minimization of the objective function tends to suppress the unconstrained entries. Therefore, constraining the scale of the first entries may promote a more reasonable estimation of $\tilde{\mathbf{c}}$, since the reverberation usually have a smaller magnitude than the direct-path and early reflections. A similar least-square problem was proposed in [30] for channel identification in the time domain. In [30], one sample of one of the two RIRs is constrained to 1. Unlike RIRs, the direct-path responses of the two CTFs are in the same tap, which allows us to constrain both of them. As shown in [53], [54], in the presence of noise interference, the estimates of CTFs obtained by respectively constraining one of the two channels are biased with different biases. Thence, a better performance is expected by constraining both channels at the same time. The solution

to (10) is

$$\hat{\tilde{\mathbf{c}}} = \frac{(\mathbf{Z}^H \mathbf{Z})^{-1} \mathbf{g}}{\mathbf{g}^\top (\mathbf{Z}^H \mathbf{Z})^{-1} \mathbf{g}} \quad (12)$$

where H denotes conjugate transpose. The estimates of $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$ are respectively $\hat{\tilde{\mathbf{a}}} = \hat{\tilde{\mathbf{c}}}_{1:\tilde{Q}}$ and $\hat{\tilde{\mathbf{b}}} = \hat{\tilde{\mathbf{c}}}_{\tilde{Q}+1:2\tilde{Q}}$.

It is obvious that $\hat{\tilde{\mathbf{a}}}$ and $\hat{\tilde{\mathbf{b}}}$ are the estimates of a normalized version of $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$. The normalization factor is $a_0 + b_0$, which however varies along the frequency bands, and leads to a gain ambiguity. To remove the gain ambiguity, we propose to further normalize the filters by the first entry of one of the filters, e.g. the first entry of $\hat{\tilde{\mathbf{a}}}$ (denoted by \hat{a}_0). Formally, the normalized filters are computed as $\hat{\tilde{\mathbf{a}}}/\hat{a}_0$ and $\hat{\tilde{\mathbf{b}}}/\hat{a}_0$, which then are the estimates of $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$ normalized by a_0 . In the k -th frequency bin, the source signal corresponding to $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$ is $s_{p,k}$, thence the source signal corresponding to $\hat{\tilde{\mathbf{a}}}/\hat{a}_0$ and $\hat{\tilde{\mathbf{b}}}/\hat{a}_0$ is $a_{0,k}s_{p,k}$. From (3), $a_{0,k}$ can be represented as

$$a_{0,k} = \sum_{n=0}^{N-1} a(n)\nu(n)e^{-j\frac{2\pi}{N}kn}, \quad (13)$$

where $\nu(n) = \sum_m \tilde{w}(m)w(m-n)$ is a window function. Therefore, $a_{0,k}|_{k=0}^{N-1}$ can be interpreted as the Fourier transform of the impulse response segment $a(n)|_{n=0}^{N-1}$ windowed by $\nu(n)$. Accordingly, the time-domain signal corresponding to $a_{0,k}s_{p,k}$ will be the convolution between $s(n)$ and $a(n)|_{n=0}^{N-1}$. The gain ambiguity is removed by consistently normalizing the filters with the early part of one channel.

This two-channel filter identification method can be extended to the multichannel case as follows. The filter vector $\tilde{\mathbf{c}}$ would stack the CTFs of all channels. As proposed in [16], [22], the signal matrix \mathbf{Z} can be organized by concatenating the signal matrices of each microphone pair. In the constraint vector \mathbf{g} , the entries corresponding to the first entry of each channel are set to 1, and the others to 0.

III. NONNEGATIVE MULTICHANNEL EQUALIZATION

Let us still consider the two-channel case. The filters $\hat{\tilde{\mathbf{a}}}/\hat{a}_0$ and $\hat{\tilde{\mathbf{b}}}/\hat{a}_0$ are the estimates of the critically sampled CTFs. We found that the complex-valued CTF estimates are not accurate enough and thus unreliable for multichannel equalization. Therefore, we only use the magnitude of CTFs to recover the STFT magnitude of source signal. To apply multichannel equalization on the oversampled microphone signals, we need to reconstruct the oversampled CTFs, which is simply done by inserting zeros between the filter coefficients. Let $\bar{\mathbf{a}} = [\bar{a}_0, \bar{a}_1, \dots, \bar{a}_{\tilde{Q}-1}]$ and $\bar{\mathbf{b}} = [\bar{b}_0, \bar{b}_1, \dots, \bar{b}_{\tilde{Q}-1}]$ denote the oversampled magnitude filters, where \tilde{Q} is the length. Note that a non-zero element appears every fourth tap, and the filter length \tilde{Q} is slightly different from the original Q due to the downsampling and upsampling operations.

Let us rewrite the microphone signals and the source signal in a vector form as $\mathbf{x} = [x_0, \dots, x_{P-1}]^\top$, $\mathbf{y} = [y_0, \dots, y_{P-1}]^\top$ and $\mathbf{s} = [s_0, \dots, s_{P-1}]^\top$, respectively. In

the previous section we assumed that the microphone signals were noise free. In this section, we explicitly introduce additive noise to the microphone signals as $\mathbf{x} = \mathbf{x}_c + \mathbf{e}_x$, where \mathbf{x}_c and \mathbf{e}_x denote the noise-free signal and noise, respectively. Similarly, define $\mathbf{y} = \mathbf{y}_c + \mathbf{e}_y$. The noise signals are assumed to be uncorrelated to the noise-free signals, to obey an i.i.d. complex Gaussian distribution, and to be spatially white. We concatenate them as $\mathbf{z}_c = [\mathbf{x}_c^\top, \mathbf{y}_c^\top]^\top$, $\mathbf{e} = [\mathbf{e}_x^\top, \mathbf{e}_y^\top]^\top$ and $\mathbf{z} = [\mathbf{x}^\top, \mathbf{y}^\top]^\top = \mathbf{z}_c + \mathbf{e}$. Note that the proposed channel identification algorithm is now directly applied to the noisy microphone signals.

From $\bar{\mathbf{a}}$, we construct the convolution matrix as

$$\bar{\mathbf{A}} = \begin{bmatrix} \bar{a}_0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \bar{a}_{\bar{Q}-1} & \bar{a}_{\bar{Q}-2} & \ddots & \bar{a}_0 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \bar{a}_{\bar{Q}-1} & \bar{a}_{\bar{Q}-2} & \cdots & \bar{a}_0 \end{bmatrix}$$

with size of $P \times P$. The matrix $\bar{\mathbf{B}}$ is defined from $\bar{\mathbf{b}}$ following the same principle. Then we concatenate the two matrices to yield $\bar{\mathbf{C}} = [\bar{\mathbf{A}}^\top, \bar{\mathbf{B}}^\top]^\top$. In this section, instead of the complex-valued CTF convolution and additive noise, we use the nonnegative approximation, namely

$$|\mathbf{z}| \approx |\mathbf{z}_c| + |\mathbf{e}| \approx \bar{\mathbf{C}}|\mathbf{s}| + |\mathbf{e}|, \quad (14)$$

where $|\cdot|$ denotes the entry-wise absolute value of a matrix or vector. Actually, based on the triangle inequality, we have $\bar{\mathbf{C}}|\mathbf{s}| \succeq |\mathbf{z}_c|$ and $|\mathbf{z}_c| + |\mathbf{e}| \succeq |\mathbf{z}|$, where \succeq denotes entry-wise vector inequality. This nonnegative approximation has been used in [43], and its noise-free case, i.e. only nonnegative CTF convolution, has also been used in [41], [42], which was shown to be a reasonable approximation.

Let $\bar{\mathbf{z}}$ and $\bar{\mathbf{s}}$ denote $|\mathbf{z}|$ and $|\mathbf{s}|$, respectively. For the noise-free case, the STFT magnitude of source signal can be recovered by solving the constrained least-square problem

$$\hat{\mathbf{s}}_{\text{ls}} = \underset{\bar{\mathbf{s}}, \text{ s.t. } \bar{\mathbf{s}} \succeq \mathbf{0}}{\operatorname{argmin}} \|\bar{\mathbf{z}} - \bar{\mathbf{C}}\bar{\mathbf{s}}\|^2. \quad (15)$$

In theory, based on the triangle inequality, the solution of (15) would be an underestimation of $\bar{\mathbf{s}}$. The underestimated magnitude is smaller than the true one, thus it will possibly suffer from some distortions, but it will not include more reverberation than $\bar{\mathbf{s}}$. In addition, as mentioned in Section II-C, the source signal corresponding to the normalized filters involves some early reflections, which provides more tolerance to the underestimation. Briefly stated, the lost information of the direct-path signal due to the underestimation could possibly be preserved in its early reflections, and vice versa.

For the noisy case, the sparsity of the speech spectra can be exploited to suppress the noise by adding an ℓ_1 -norm minimization on the source signal to (15). For the magnitude vector $\bar{\mathbf{s}}$, the ℓ_1 -norm is actually the element-summation, i.e. $\mathbf{1}^\top \bar{\mathbf{s}}$, where $\mathbf{1}$ is a vector with all entries equal to 1. In the spirit of [37], we realize the ℓ_1 -norm minimization by solving

the constrained optimization problem:

$$\begin{aligned} \hat{\mathbf{s}}_{\ell_1} &= \underset{\bar{\mathbf{s}}}{\operatorname{argmin}} \mathbf{1}^\top \bar{\mathbf{s}} \\ \text{s.t. } &\bar{\mathbf{s}} \succeq \mathbf{0}, \|\bar{\mathbf{z}} - \bar{\mathbf{C}}\bar{\mathbf{s}}\|^2 \leq \delta. \end{aligned} \quad (16)$$

The ℓ_2 -norm fitting cost $\|\bar{\mathbf{z}} - \bar{\mathbf{C}}\bar{\mathbf{s}}\|^2$ is relaxed to at most δ .

The relaxing tolerance δ is related to the noise power in the microphone signals. The magnitude convolution $\bar{\mathbf{C}}\bar{\mathbf{s}}$ should target the magnitude of the noise-free microphone signal, i.e. $|\mathbf{z}_c|$. Therefore, the ℓ_2 -norm fitting cost should be relaxed to $\|\bar{\mathbf{z}} - |\mathbf{z}_c|\|^2$. Based on the approximation $\bar{\mathbf{z}} - |\mathbf{z}_c| \approx |\mathbf{e}|$, we can set the tolerance using the power of the complex-valued noise, i.e. $\|\mathbf{e}\|^2$. However, note that $\|\mathbf{e}\|^2$ is actually an overestimation of $\|\bar{\mathbf{z}} - |\mathbf{z}_c|\|^2$. Let σ_{ex}^2 and σ_{ey}^2 denote the noise PSD in the two channels, which can be estimated from the pure noise signal for stationary noise, or estimated by a noise PSD estimator for non-stationary noise, e.g. [55]. Then $\|\mathbf{e}_x\|^2$ (resp. $\|\mathbf{e}_y\|^2$) follow an Erlang distribution with mean $P\sigma_{ex}^2$ ($P\sigma_{ey}^2$) and variance $P\sigma_{ex}^4$ ($P\sigma_{ey}^4$). For spatially white noise, $\|\mathbf{e}\|^2$ has the mean $P(\sigma_{ex}^2 + \sigma_{ey}^2)$ and variance $P(\sigma_{ex}^4 + \sigma_{ey}^4)$. The tolerance with respect to noise is set to

$$\delta_e = P(\sigma_{ex}^2 + \sigma_{ey}^2) - 2\sqrt{P(\sigma_{ex}^4 + \sigma_{ey}^4)}. \quad (17)$$

Subtracting two times the standard deviation makes the probability that the ℓ_2 -norm fitting cost being larger than $\|\mathbf{e}\|^2$ very small. When the ℓ_2 -norm fitting cost is allowed to be larger than $\|\mathbf{e}\|^2$, the minimization of $\mathbf{1}^\top \bar{\mathbf{s}}$ can distort the source signal. As a result, some noise remains in the estimated STFT magnitude of source signal. In addition, this mitigates the inaccuracy of the assumption $\bar{\mathbf{z}} - |\mathbf{z}_c| \approx |\mathbf{e}|$.

Besides, the ℓ_2 -norm fitting cost should also be relaxed to take into account the filter estimation error and the fact that the nonnegative convolution $\bar{\mathbf{C}}\bar{\mathbf{s}}$ does not accurately fit $\bar{\mathbf{z}}_c$ by definition. The inaccuracy is akin to the level of the noise-free signal, i.e. $\lambda_c = \|\mathbf{z}_c\|^2$, which can be estimated by spectral subtraction as $\hat{\lambda}_c = \max(\|\mathbf{z}\|^2 - P(\sigma_{ex}^2 + \sigma_{ey}^2), 0)$. Empirically, the tolerance with respect to the noise-free signal is set to $\delta_c = 0.05\hat{\lambda}_c$. The relaxing tolerance can be set to $\delta_e + \delta_c$. However, for this quantity, the ℓ_2 -norm constraint in (16) is not definitely feasible, since both δ_e and δ_c are set to be relatively small to avoid the source signal distortion. This often happens when the noise power is very low. The minimum ℓ_2 -norm fitting error is defined in (15), and is computed as $\|\bar{\mathbf{C}}\hat{\mathbf{s}}_{\text{ls}} - \bar{\mathbf{z}}\|^2$. Overall, taking this error as the lower bound, the relaxing tolerance is set as

$$\delta = \max(\delta_e + \delta_c, 1.05 \|\bar{\mathbf{C}}\hat{\mathbf{s}}_{\text{ls}} - \bar{\mathbf{z}}\|^2), \quad (18)$$

where 1.05 is a slack factor.

We need to solve (15) to determine δ , and solve (16) to recover the STFT magnitude of the source signal. Both of them are convex optimization problems with an inequality constraint. We adopt the PDIPM method [52] to solve them. The PDIMP algorithm is briefly presented in the Appendix. The multichannel extension of this multichannel equalization method is straightforward. The filter matrix $\bar{\mathbf{C}}$ and signal vector $\bar{\mathbf{z}}$ are constructed by stacking all the channels.

At each frequency bin k , the vector $\hat{\mathbf{s}}_{\ell_1}$ contains the sequence of estimated source signal magnitudes $\bar{s}_{p,k}$. The phase of one of the microphone signals is taken as the corresponding phase and we thus have $\hat{s}_{p,k} = \bar{s}_{p,k} e^{j \arg[x_{p,k}]}$, where $\arg[\cdot]$ is the phase of complex number. The time-domain source signal $\hat{s}(n)$ is obtained by applying the inverse STFT to $\hat{s}_{p,k}$. Note that the proposed method outputs a single-channel dereverberated source signal. As mentioned in Section II-C, the time-domain signal $\hat{s}(n)$ is an estimation of $s(n) \star a(n)|_{n=0}^{N-1}$, where $a(n)$ starts with the direct-path impulse response. The window size N is generally significantly larger than the duration of the direct-path impulse response, thus $\hat{s}(n)$ also includes the early reflections, e.g. 64 ms in this work.

IV. EXPERIMENTS WITH SIMULATED BINAURAL DATA

In this section, we present a series of experiments with simulated (two-channel) binaural data. A set of binaural room impulse responses (BRIRs) were generated with the ROOMSIM simulator [56] combined with the head-related impulse responses (HRIRs) of the KEMAR dummy head [57]. For the KEMAR dummy head, the pre-measured HRIRs for a large set of discrete directions (for both azimuth and elevation) are available. To simulate the filtering of a reflection coming from a given direction, the pre-measured HRIR of discrete direction closest to the reflection direction is used as an approximation of the HRIR of the reflection direction. This procedure is automatically applied in the ROOMSIM simulator [56]. The simulated room is of dimension 5 m \times 8 m \times 3 m. The dummy head is located at (1 m, 4 m, 1.5 m). Sound sources are placed in front of the dummy head with azimuths (relative to the dummy head center) varying from -90° to 90° , spaced by 5° , and an elevation of 0° . The head-to-source distances were always 2 m. Two reverberation times, i.e. $T_{60} = 0.5$ s and 0.79 s, are simulated by adjusting the absorption coefficients of the walls. Speech signals from the TIMIT dataset [58] are taken as the source signals, with a duration of about 4 s. To generate the noisy microphone signals, a spatially uncorrelated stationary speech-like noise is added with signal-to-noise ratio (SNR) of 0, 5, 10, 15, 20 dB, respectively. For each acoustic condition, 50 runs are performed with random directions and speech utterances.

The sampling rate is 16 kHz. As already mentioned, the STFT uses a Hamming window with $N = 1,024$ (64 ms) and $L = N/4 = 256$ (16 ms). The noise PSD is estimated from the pure noise signals for various SNRs. The CTF length Q (or \bar{Q}) is related to the reverberation time, and is the only prior knowledge that the proposed method requires. The setting of Q influences the performance to a large extent. If Q is too small, the CTFs can not model the real RIRs, and will be inaccurately estimated. If Q is too large, the CTFs will be identified with a long noisy tail. Thence, the CTF length should be set to cover the major part of the RIRs, and also to avoid a heavy tail. Based on some pilot experiments, the CTF length is set to approximately 0.5 times T_{60} , e.g. 256 ms for $T_{60} = 0.5$ s and 384 ms for $T_{60} = 0.79$ s, correspondingly, Q (\bar{Q}) is 16

and 24 (4 and 6). An example of CTF identification is shown in Fig. 2, which demonstrates that this is a reasonable choice.

Three STFT-based baseline methods are compared. **i)** In [43], a nonnegative representation similar to (14) is used. The magnitudes of CTF, source spectra and noise spectra are iteratively updated using the multiplicative update method. We refer to this method as nonnegative iterative method (NIM). A decaying structure is imposed on the CTF magnitude, not for each frequency separately, but for the summation of CTF squared magnitude over frequencies. This decaying structure can partially overcome the local optima problem. The software provided by the authors of [43] is used. The same STFT configuration and CTF length are used as the proposed method, which performs the best in our experiments. This method is a single-channel method, thus we will only illustrate an example of this method, in terms of both CTF identification and source spectra estimation. In addition, we also test the dereverberated signals obtained using the multichannel CTF magnitude estimates of NIM and the proposed nonnegative multichannel equalization method. We refer to this method as NIM-NME (nonnegative multichannel equalization). The CTF magnitude is individually estimated for each channel, thence the multichannel estimates have an uncertain scale misalignment. To remove this misalignment, we adjust the scale of the multichannel estimates according to the scale of the theoretical CTF computed by (3) based on the true time-domain filter. The comparison between NIM-NME and the proposed method is relatively fair when the number of channels is not large, e.g. the two-channel case will be considered in the following experiments. For quantitative comparison, the CTF estimates are downsampled as is done in this work to also involve the 64 ms early reflections in the dereverberated signal. **ii)** The WPE method [7], [11]. In our experiments, a spectral subtraction method is applied to the single-channel WPE output for denoising, in the spirit of [11]. For spectral subtraction, the noise PSD is estimated using the single channel noise estimator [55], and an advanced speech estimator, i.e. optimally-modified log-spectral amplitude [59], is used. The software provided by the authors of [7] is used for WPE. The STFT configuration is set as the default values in the software, namely using a Hanning window with the length of 512 and shift size of 128. Under the conditions with $T_{60} = 0.5$ s and 0.79 s, the number of filter coefficients is set to 50 and 80, respectively, which correspond to about $0.8T_{60}$. The prediction delay is set to 8 to involve the 64 ms early reflections in the dereverberated signal. The first channel is taken as the target channel. **iii)** The CDR method [5]. The software provided by the authors of [5] and the *Proposed 2* estimator therein is used. The true direction of arrival (DOA) is adopted. The microphone distance is set to 18 cm according to the size of KEMAR head. For all the three methods, other parameters not mentioned are kept at their default values.

Three metrics are used to quantitatively evaluate the dereverberation performance, i) a non-intrusive metric, normalized speech-to-reverberation modulation energy ratio (SRMR) [60], and two intrusive metrics ii) perceptual evaluation of speech quality (PESQ) [61] and iii) log-spectral distance (LSD)

[13] with the desired dynamic range of 50 dB. Some early reflections are preserved in all the four methods, specifically 64 ms in WPE, NIM-NME and the proposed method, while an unknown amount in CDR. The reference signal used to measure PESQ and LSD is set as the early (64 ms) reverberated signal, which is generated by convolving the source signal with the first 64 ms (starting from the direct-path) of the first-channel RIR. The CDR method in [5] uses a 50 ms early reverberated signal as the reference, however the use of 64 ms in our experiments does not lead to a significant difference. The dereverberated signal of various methods is an estimation of the reference signal up to a time shift and/or a gain factor. Therefore, as is for PESQ, the signals are aligned and amplitude normalized prior to the computation of LSD. For SRMR and PESQ, the higher the better, and for LSD, the lower the better. Note that the average scores over the 50 runs are reported.

A. An Example of CTF Identification

Fig. 2 (a) depicts an example of CTF identification obtained with the proposed method. The theoretical CTFs computed by (3) with $k = k'$ is taken as a baseline. For comparison, the estimate of the oversampled CTF is also shown, which is computed by constructing the convolution matrix \mathbf{X} without downsampling the microphone signals. The magnitude of downsampled CTFs globally follow the curve of the theoretical CTFs, which indicates the accuracy of CTF identifications. By contrast, the magnitude of the CTFs estimated without downsampling deviates from the theoretical CTFs. The downsampled CTF magnitudes estimated from the noise-free and noisy microphone signals have only a small difference, which indicates that the CTF identification is not significantly degraded by the spatially white microphone noise. Fig. 2 (b) depicts the CTF magnitudes obtained with NIM. Note that the method is separately applied to the two channels. We can see that the CTF magnitudes estimated from the noise-free and noisy microphone signals are very similar, which indicates the accuracy of the nonnegative model and the efficiency of the multiplicative update method. The CTF magnitudes are well estimated for the high magnitude frames, while are underestimated for the low magnitude frames. In addition, the reverberation frames are underestimated relative to the direct-path frame. These are possibly due to the local optima problem.

B. Spectrogram Examples

Fig. 3 depicts the spectrogram examples of the proposed method for both noise-free and noisy signal. As mentioned in the methodology part, the identified CTFs suffer from some errors i) the CTF approximation error, namely the loss of cross-band information, ii) the frequency aliasing of CTF caused by the critical sampling, iii) the magnitude approximation error, namely the loss of phase information. To evaluate the influence of these errors, we also show the dereverberated signals obtained using i) the theoretical CTF computed by (3) based on the true time-domain filter, with

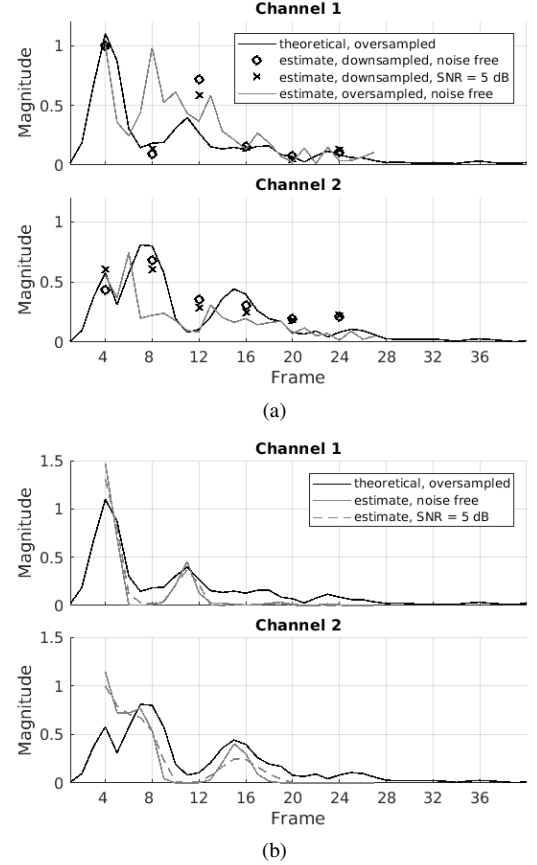


Fig. 2: Example of CTF identification at frequency 1,900 Hz, obtained with (a) the proposed method and (b) the nonnegative iterative method [43]. $T_{60}=0.79$ s.

the frame step $L = N/4$. The multichannel equalization is carried out by minimizing the regularized ℓ_2 -norm cost function $\|\mathbf{z} - \mathbf{C}\mathbf{s}\| + \lambda\|\mathbf{s}\|_1$, where $\|\cdot\|_1$ denotes ℓ_1 -norm. Note that all terms are complex-valued. This optimization method was proposed in [51] for multiple sources, and is able to handle the single-source case, ii) the magnitude of theoretical CTF and the proposed nonnegative multichannel equalization method. In addition, to demonstrate the use of critically sampled CTF, we also present the dereverberated signals obtained using iii) the identified oversampled CTF and the optimization method in [51], iv) the magnitude of identified oversampled CTF and the proposed nonnegative multichannel equalization. Fig. 3 also depicts the spectrogram examples for these four cases for the noise-free signal. Fig. 3 (a), (b) and (c) respectively illustrate the early reverberated signal, noise-free and noisy microphone signal. The smearing effect of reverberation is clearly seen in the microphone signal.

The theoretical CTF do not have the gain ambiguity across frequencies, thus the frequency-wise dereverberated signals are consistent with the source signal. It can be observed from Fig. 3 (d) that the dereverberated signal does not include early reflections, which can also be verified by listening to it. However, we can perceive a small delayed replica of the original source signal, which is not obvious in the spectrogram, and is possibly caused by the loss of cross-band

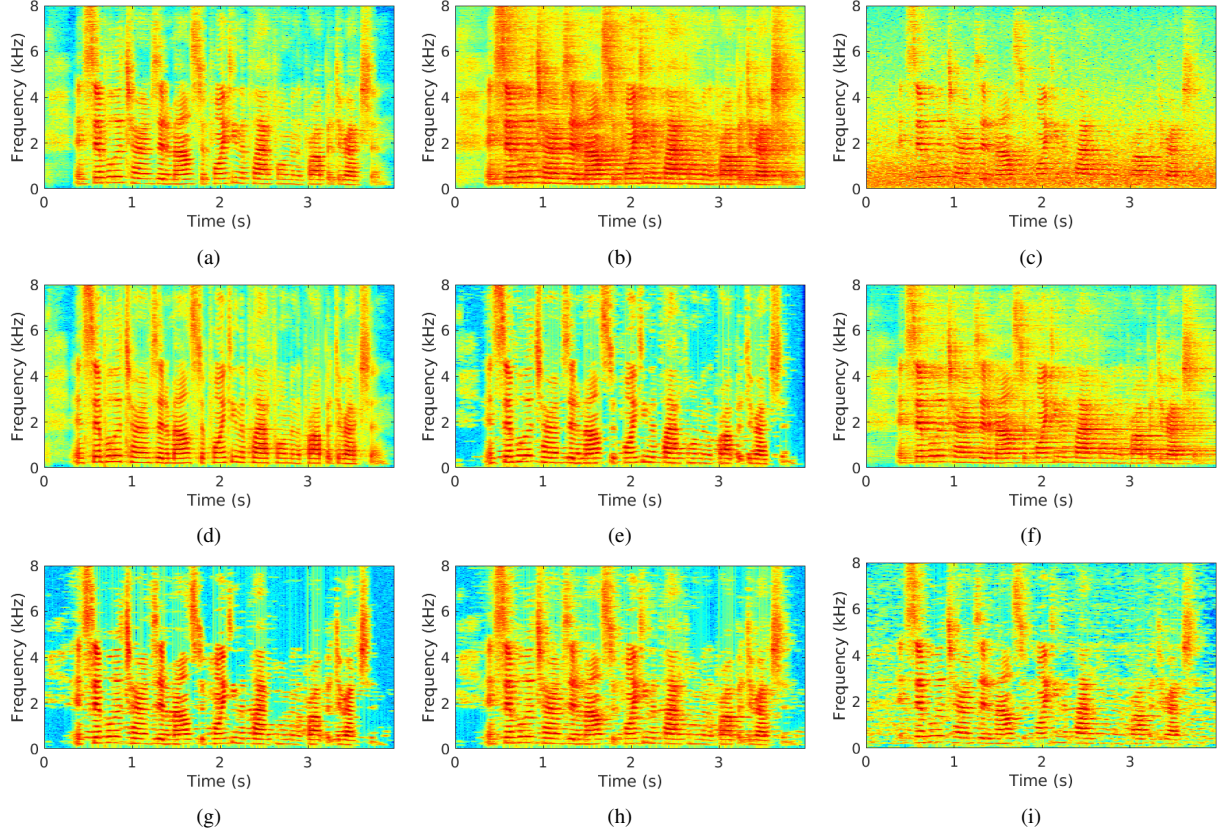


Fig. 3: Spectrogram examples. $T_{60} = 0.79$ s. (a), (b) and (c) are respectively the early reverberated signal, noise-free and noisy microphone signal. (d)~(h) are the outputs for the noise-free signal obtained by using (d) theoretical oversampled CTF, (e) theoretical oversampled CTF magnitude, (f) identified oversampled CTF, (g) identified oversampled CTF magnitude and (h) the proposed identified critically sampled CTF magnitude. (i) is the output obtained using the proposed identified critically sampled CTF magnitude for the noisy signal.

information. In Fig. 3 (e), it can be seen that the source signal is also recovered by using the magnitude of theoretical CTF. Compared with Fig. 3 (d), some spectral distortions exist, especially in the low power regions, due to the underestimation of \bar{s} , but reverberation is clearly removed. This confirms that the underestimated signal only suffers from distortions, but not from more reverberation. A small replica still exists, and is perceptually less natural than the signal in Fig. 3 (d). In Fig. 3 (f), the identified oversampled CTF does not efficiently dereverberate the microphone signal. When using its magnitude (Fig. 3 (g)), the main structure of the source signal is recovered. However, there are some distortions, specifically, many spectral regions before a strong harmonic are wrongly enhanced, for example the region around 0.5 s at 4 kHz. This is due to the identification inaccuracy of the oversampled CTF (magnitude) as shown in Fig. 2 (a).

From Fig. 3 (h), it can be seen that the identified critically sampled CTF achieves a good estimation of the early reverberated signal. There also exist some spectral distortions, especially in the low power regions. Compared to Fig. 3 (d) and (e), the noise/distortions sounds weaker, since the early reflections enhance the desired signal. This confirms that the early reflections provide more tolerance to the underestimation of \bar{s} . However, the distortions sound unnatural, like musical noise. Fig. 3 (i) illustrates the dereverberated signal for the

noisy microphone signal. As mentioned in Section III, an approximated noise magnitude model, i.e. $|e| \approx |z| - |z_c|$, is used, which leads to an overestimation of noise power and thus a sparser source signal estimate. However, we can observe from Fig. 3 (i) that the source signal is not overly sparsed, which indicates that this problem is not severe. Some residual noise is present in Fig. 3 (i), which masks the low power speech spectra. Informal listening tests show that the small replica and musical noise that were present in the noise-free case are not clearly audible for this noisy case, since it is overshadowed by the residual noise.

Fig. 4 depicts the spectrogram examples for the comparison methods. For the noise-free case, NIM (Fig. 4 (a)) enhances the direct-path and some early reflections. However, some reverberation remains due to the underestimation of the CTF magnitude of the reverberation frames. By combining with the proposed nonnegative multichannel equalization method, NIM-NME (Fig. 4 (b)) has less speech distortion than NIM, especially in the high frequency region. Compared with the proposed method, more reverberation remains by NIM-NME. This confirms that the proposed method achieves a better CTF magnitude estimation than NIM. WPE (Fig. 4 (c)) achieves a very good estimate of the early reverberated signal in terms of high reverberation suppression and low desired signal distortion. CDR (Fig. 4 (d)) preserves the desired direct-

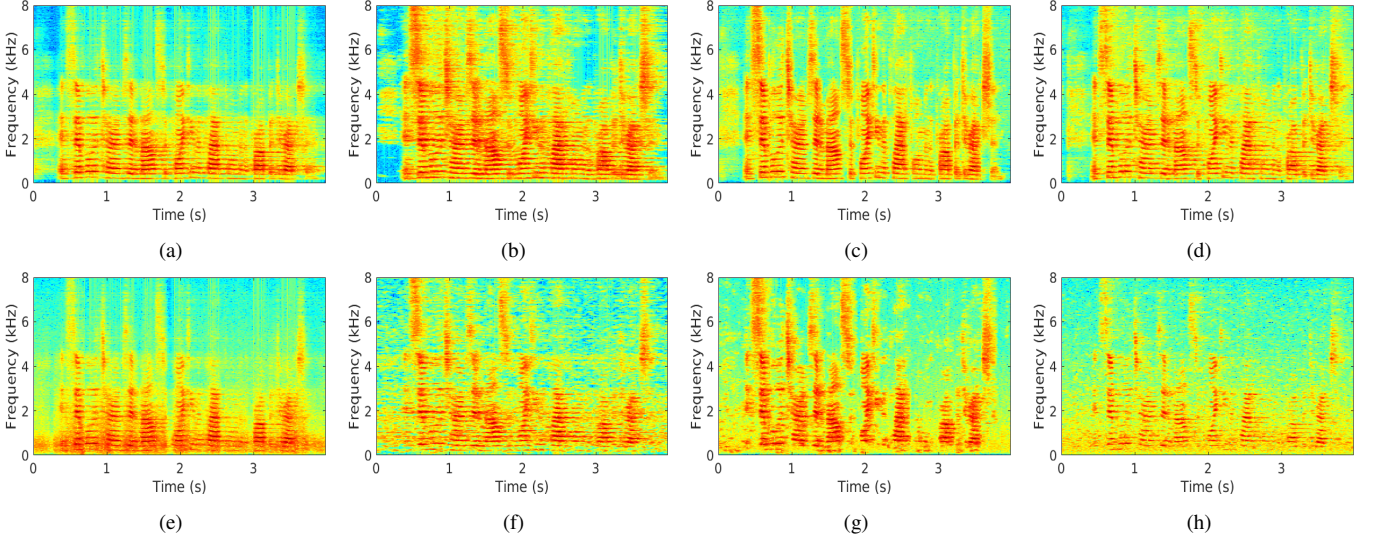


Fig. 4: Spectrogram examples of the comparison methods. Please see Fig. 3 for the microphone signals. (a), (b), (c) and (d) are the outputs for noise-free signal of (a) NIM (the output of first channel), (b) NIM-NME, (c) WPE and (d) CDR, respectively. (e), (f), (g) and (h): the same for the noisy signal.

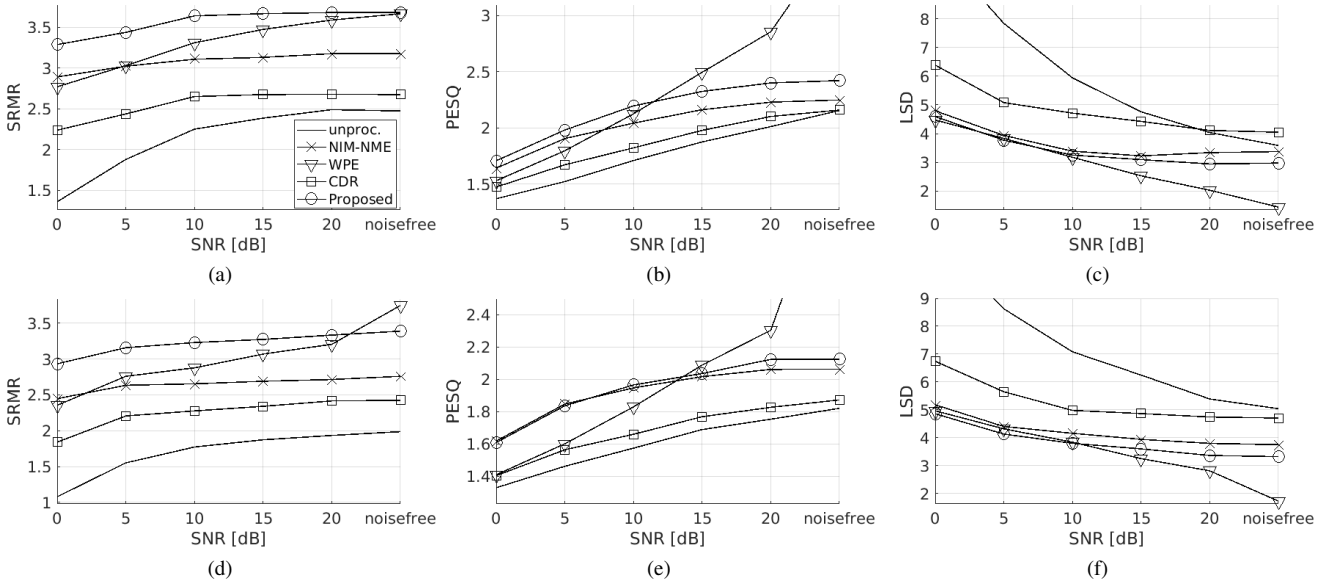


Fig. 5: Dereverberation performance as a function of input SNR. ‘unproc.’ denotes the score of the unprocessed microphone signal. (a), (b) and (c) are for $T_{60} = 0.5$ s, (d), (e) and (f) are for $T_{60} = 0.79$ s.

path and some early reflections, but some amount of late reverberation remains. For the noisy case, NIM (Fig. 4 (e)) and NIM-NME (Fig. 4 (f)) remove most of the noise, and recover a similar speech spectra as in the noise-free case that some reverberation remains. For WPE (Fig. 4 (g)), the noise and reverberation are well reduced, while more speech distortion present than the noise-free case due to the spectral subtraction. CDR (Fig. 4 (h)) reduces the microphone noise to a certain extent. Informal listening tests show that the residual reverberation is similar to the noise-free case.

C. Quantitative Dereverberation Performance

Fig. 5 shows the quantitative results of the proposed method, namely using the identified critically sampled CTF magnitude and nonnegative multichannel equalization method, and three comparison methods, i.e. NIM-NME, WPE and CDR. For all the four methods, it is not surprising that all the performance measures for the case of $T_{60} = 0.79$ s are worse than the measures for the case of $T_{60} = 0.5$ s. In terms of SRMR, which mainly measures the amount of reverberation and noise, the proposed method and WPE achieve a comparable score for the high SNR cases (larger than 15 dB). The score of WPE decreases faster than the proposed method with the decrease

of SNR. The SRMR scores of NIM-NME are considerably lower than the scores of the proposed method for all the SNR conditions, which is consistent with spectrogram examples shown in Fig. 4 that more reverberation remains by NIM-NME. The high SRMR score indicates that the proposed method can efficiently suppress reverberation and noise. PESQ and LSD measure the difference between the enhanced signal and the reference signal. In terms of PESQ and LSD, compared with WPE, the proposed method performs worse for the high SNR cases (approximately larger than 10 dB), whereas it performs better for the low SNR cases. Compared with NIM-NME, the PESQ and LSD scores of the proposed method are larger for the high SNR cases, while are comparable for the low SNR cases. The proposed method underestimates some speech spectra due to the nonnegative approximation, which makes a relatively large difference between the enhanced signal and the reference signal. The underestimation of a high power spectra is not often audible, whereas the distortion of a low power spectra can sound like noise. With the decrease of SNR, WPE also has a larger spectral distortion due to the inaccuracy of linear prediction and spectral subtraction. The CTF magnitude estimation of NIM is robust against noise, but is less accurate than the proposed method. We remind that NIM is a single channel method, namely without using the spatial information. The CDR method achieves the lowest scores. Overall, the performance of all the four methods degrade with the decrease of SNR, due to a larger estimation error of the filters/parameters and more residual noise. The proposed method, NIM-NME and CDR have a similar performance degradation rate, while WPE has the largest one.

V. EXPERIMENTS WITH MULTICHANNEL DATA

To evaluate the proposed method with multiple microphones, and with real impulse responses or real recordings, we conducted two sets of experiments with different datasets.

The *multichannel impulse response dataset* [62] was measured using a 8-channel linear microphone array in the speech and acoustic lab of Bar Ilan University. The reverberation time is controlled by 60 panels covering the room facets. The configuration of the impulse response dataset used in this experiment is i) the reverberation time T_{60} is 0.61 s, ii) the microphone-to-source distance is 2 m, iii) the source direction is in $[-90^\circ, 90^\circ]$, iv) the number of microphone is 2 or 4 (the central two or four microphones). Again, TIMIT signals and spatially white stationary speech-like noise are taken as the source signal and additive noise, respectively. Two input SNRs, i.e. 20 and 5 dB, are tested. The parameters are set as for experiments with the binaural dataset in Section IV, except that, according to the reverberation time of 0.61 s, the CTF length is set to 320 ms (20 taps) for the proposed method and NIM-NME, and the filter length for WPE is set to 60 and 20 for the 2-channel and 4-channel cases, respectively. CDR is only applicable for the 2-channel case, and the corresponding microphone spacing, i.e. 8 cm, is used.

TABLE I: The results for the multichannel impulse response dataset.

	noise	unproc.	NIM-NME	CDR	WPE		Proposed	
			2-ch	2-ch	2-ch	4-ch	2-ch	4-ch
SRMR	20 dB	2.54	3.20	2.81	3.30	3.37	3.40	3.42
	5 dB	1.96	3.07	2.68	2.93	3.06	3.24	3.25
PESQ	20 dB	2.37	2.71	2.48	3.13	3.31	2.70	2.81
	5 dB	1.60	2.08	1.78	1.92	2.00	2.04	2.24
LSD	20 dB	3.84	2.89	3.54	1.67	1.55	2.46	2.35
	5 dB	7.16	3.33	4.43	3.23	3.15	3.23	3.05

TABLE II: The SRMR scores for the REVERB challenge dataset.

	dis	unproc.	WPE		Proposed	
			2-ch	8-ch	2-ch	8-ch
SRMR	<i>near</i>	2.07	2.98	3.16	3.20	3.20
	<i>far</i>	1.90	2.81	2.97	3.03	3.07

The results are shown in Table I. For the 2-channel case, it can be seen that the performance measures of all the four methods are almost consistent with the results obtained on the binaural data. As expected, the proposed method and WPE achieve better scores with 4 channels than with 2 channels. For the proposed channel identification method, the identification of each channel is carried out by using the cross-relations with all the other channels, thence a more robust identification can be achieved by increasing the number of channels. For the proposed multichannel equalization method, a larger number of channels will give a larger data size for the ℓ_2 -norm fitting problem in (15) and (16), which leads to a smaller error covariance of the least square estimation. In addition, informal listening tests show that the musical noise presented in the 2-channel case is noticeably suppressed in the 4-channel case.

The *REVERB challenge RealData* [63] was recorded in a room with T_{60} of 0.7 s. It contains 2 types of microphone-to-speaker distances, namely *near* (1 m) and *far* (2.5 m), which respectively have 90 and 89 recordings with different directions. We use the 2-channel and 8-channel RealData for development (dev). According to T_{60} , the CTF length is also set to 320 ms (20 taps) for the proposed method, and the filter length for WPE is set to 70 and 10 for the 2-channel and 8-channel cases, respectively. Since the pure noise signal is not available for this dataset, the noise PSD is estimated using the single channel noise estimator [55]. For NIM-NME, to align the CTF magnitude individually estimated for each channel, the prior knowledge of the spatial relation between multiple channels, e.g. the theoretical CTFs used in the previous experiments, is required, which is not available for this dataset, thus NIM-NME is not tested. CDR needs the prior knowledge of either the DOA or the noise coherence, or both, which are not available for this dataset, thus CDR is also not tested. Only the SRMR score is given due to the lack of reference signal. The results are shown in Table II. The SRMR scores of the *near* case are higher, since the *near* case has a larger direct-to-reverberation ratio than the *far* case, in other words, the desired direct-path signal (and early reflections) is less contaminated by the late reverberation.

Audio examples for all experiments presented in this paper are available in our website.²

²<https://team.inria.fr/perception/research/ctf-dereverberation>

TABLE III: The specifications of the optimization problems (15) and (16), where $\mathbf{I} \in \mathbb{R}^{P \times P}$ denotes the identity matrix, $\mathbf{0}_{P \times P} \in \mathbb{R}^{P \times P}$ denotes the matrix with all entries equal to 0.

	$f_0(\bar{\mathbf{s}})$	$f_i(\bar{\mathbf{s}})$	λ	$\nabla f_0(\bar{\mathbf{s}})$	$\nabla^2 f_0(\bar{\mathbf{s}})$	$Df(\bar{\mathbf{s}})$	$\nabla^2 f_i(\bar{\mathbf{s}})$
(15)	$\ \bar{\mathbf{C}}\bar{\mathbf{s}} - \bar{\mathbf{z}}\ ^2$	$-\bar{s}_i, i \in [1, P]$	$\lambda_i, i \in [1, P]$	$\bar{\mathbf{C}}^\top (\bar{\mathbf{C}}\bar{\mathbf{s}} - \bar{\mathbf{z}})$	$\bar{\mathbf{C}}^\top \bar{\mathbf{C}}$	$-\mathbf{I}$	$\mathbf{0}_{P \times P}, \forall i$
(16)	$\mathbf{1}^\top \bar{\mathbf{s}}$	$-\bar{s}_i, i \in [1, P];$ $\ \bar{\mathbf{C}}\bar{\mathbf{s}} - \bar{\mathbf{z}}\ ^2 - \delta, i = P + 1$	$\lambda_i, i \in [1, P];$ λ_{P+1}	$\mathbf{1}$	$\mathbf{0}_{P \times P}$	$[-\mathbf{I}, \bar{\mathbf{C}}^\top (\bar{\mathbf{C}}\bar{\mathbf{s}} - \bar{\mathbf{z}})]^\top$	$\mathbf{0}_{P \times P}, i \in [1, P];$ $\bar{\mathbf{C}}^\top \bar{\mathbf{C}}, i = P + 1$

VI. CONCLUSIONS

In this paper, a blind multichannel speech dereverberation and noise reduction method has been proposed. The cross-relation method was extended to the STFT domain to circumvent the problem of near-common zeros for long channel filters. The common zeros caused by the oversampling of STFT is solved by forcing the channel filters to be critically sampled. A constrained least-square problem was used to estimate the CTFs, which is robust to the noise interference and the filter length determination error. The CTF-based multichannel equalization is then proposed in the magnitude domain. The sparsity of the source signal is exploited. An optimization problem with respect to the ℓ_1 -norm of the STFT magnitude of source signal and the ℓ_2 -norm fitting cost between the STFT magnitude of microphone signals and the nonnegative image source signal was proposed to reduce the microphone noise and the influence of filter perturbations. A series of experiments have been carried out. It is confirmed that the identified CTF magnitude is reliable, even for the high reverberant case, and is robust to the spatially white noise, even for the low SNR case. In the nonnegative multichannel equalization method, the tolerance setting scheme for the ℓ_2 -norm fitting cost works well for noise reduction.

Overall, this paper proposes a multichannel CTF (magnitude) identification approach and a nonnegative multichannel equalization approach, and thus a practical blind dereverberation and noise reduction method in the family of multichannel (nonnegative) equalization technique.

APPENDIX

PRIMAL-DUAL INTERIOR-POINT METHOD

The book [52] provides a general optimization algorithm for a convex objective function $f_0(x)$ with a set of inequality constraints of the form $f_i(x) \leq 0, i \in [1, m]$ and an affine equality constraint. Here x denotes the optimization variable and m is the number of inequality constraints. Note that there is no affine equality constraint in the presented problems. Define the vector $f(x) = [f_1(x), \dots, f_m(x)]^\top$ including all the inequality functions, and its derivative matrix $Df(x) = [\nabla f_1(x), \dots, \nabla f_m(x)]^\top$, where ∇ denotes gradient operator. Let λ_i denote the dual variable corresponding to the inequality constraint $f_i(x) \leq 0$. The dual variable vector is $\lambda = [\lambda_1, \dots, \lambda_m]$. In PDIPM, the inequality constraint is approximately formulated as an equality constraint by the logarithmic barrier function. The parameter t sets the accuracy of the logarithmic barrier approximation, the larger t , the better

the approximation. The PDIPM is summarized in Algorithm 1, with variable update in Step 3 given by:

$$\begin{bmatrix} \Delta x \\ \Delta \lambda \end{bmatrix} = - \begin{bmatrix} \nabla^2 f_0(x) + \sum_i \lambda_i \nabla^2 f_i(x) & Df(x)^\top \\ -\text{diag}(\lambda) Df(x) & -\text{diag}(f(x)) \end{bmatrix}^{-1} \times \begin{bmatrix} \nabla f_0(x) + Df(x)^\top \lambda \\ -\text{diag}(\lambda) f(x) - (1/t)\mathbf{1} \end{bmatrix}. \quad (19)$$

In Algorithm 1, the so-called surrogate duality gap $\hat{\eta}^{(n)}$ is decreasing with the iterations, thence the parameter t is increased by the factor μ (a positive value of the order of 10). The goal of the line search (Step 2) is to find the largest step-length $\zeta^{(n)}$. The convergence criterion is set to guarantee a high optimization and the feasibility of the variables. We refer to [52] for more details. To apply the PDIPM to the problems (15) and (16), the general quantities in Algorithm 1 should be accordingly specified. Table III gives the specifications for both (15) and (16). For solving (15), a good initialization is to set $\bar{\mathbf{s}}^{(0)} = |\mathbf{x}|$ and $\lambda^{(0)}$ is set to an arbitrary positive vector ($10 \cdot \mathbf{1}$ in this paper). For solving (16), a good initialization is to set $\mathbf{s}^{(0)}$ and $\lambda^{(0)}$ as the solution of (15).

Algorithm 1 Primal-dual interior-point method

Iteration step $n = 0$.

repeat

- 1 Compute $\hat{\eta}^{(n)} = -f(x)^\top \lambda$, Set $t^{(n)} := \mu m / \hat{\eta}^{(n)}$,
- 2 Line search the step-length $\zeta^{(n)}$,
- 3 Update variables $x^{(n+1)} = x^{(n)} + \zeta^{(n)} \Delta x^{(n)}$,
and $\lambda^{(n+1)} = \lambda^{(n)} + \zeta^{(n)} \Delta \lambda^{(n)}$.

until Convergence.

REFERENCES

- [1] I. Arweiler and J. M. Buchholz, "The influence of spectral characteristics of early reflections on speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 130, no. 2, pp. 996–1005, 2011.
- [2] E. A. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, 2009.
- [3] S. Braun and E. A. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, pp. 1–5, 2013.
- [4] A. Kuklański, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1595–1608, 2016.
- [5] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006–1018, 2015.
- [6] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 85–88, 2008.

- [7] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.
- [8] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1509–1520, 2015.
- [9] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 231–246, 2009.
- [10] N. Ito, S. Araki, and T. Nakatani, "Probabilistic integration of diffuse noise suppression and dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5167–5171, 2014.
- [11] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, et al., "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge," in *REVERB Workshop*, 2014.
- [12] O. Schwartz, S. Gannot, E. Habets, et al., "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, 2015.
- [13] O. Schwartz, S. Gannot, and E. Habets, "An expectation-maximization algorithm for multi-microphone speech dereverberation and noise reduction with coherence matrix estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1495–1510, 2016.
- [14] B. Schwartz, S. Gannot, and E. A. Habets, "An online dereverberation algorithm for hearing aids with binaural cues preservation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–5, 2015.
- [15] B. Schwartz, S. Gannot, and E. A. Habets, "Online speech dereverberation using kalman filter and EM algorithm," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 394–406, 2015.
- [16] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on signal processing*, vol. 43, no. 12, pp. 2982–2993, 1995.
- [17] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 2, pp. 145–152, 1988.
- [18] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1–12, 2007.
- [19] F. Lim, W. Zhang, E. A. Habets, and P. A. Naylor, "Robust multichannel dereverberation using relaxed multichannel least squares," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1379–1390, 2014.
- [20] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 680–693, 2016.
- [21] E. Moulines, P. Duhamel, J.-F. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters," *IEEE Transactions on signal processing*, vol. 43, no. 2, pp. 516–525, 1995.
- [22] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 1074–1090, 2003.
- [23] A. P. Liavas, P. A. Regalia, and J.-P. Delmas, "Robustness of least-squares and subspace methods for blind channel identification/equalization with respect to effective channel undermodeling/overmodeling," *IEEE transactions on signal processing*, vol. 47, no. 6, pp. 1636–1645, 1999.
- [24] A. P. Liavas, P. A. Regalia, and J.-P. Delmas, "Blind channel approximation: effective channel order determination," *IEEE Transactions on Signal Processing*, vol. 47, no. 12, pp. 3336–3344, 1999.
- [25] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 11–24, 2003.
- [26] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 882–895, 2005.
- [27] X. Lin, N. D. Gaubitch, and P. A. Naylor, "Two-stage blind identification of SIMO systems with common zeros," in *14th European Signal Processing Conference*, pp. 1–5, 2006.
- [28] P. A. Naylor, X. Lin, and A. W. Khong, "Near-common zeros in blind identification of simo acoustic systems," in *IEEE Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, pp. 21–24, 2008.
- [29] X. Lin, A. W. Khong, and P. A. Naylor, "A forced spectral diversity algorithm for speech dereverberation in the presence of near-common zeros," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 888–899, 2012.
- [30] Y. Lin, J. Chen, Y. Kim, and D. D. Lee, "Blind channel identification for speech dereverberation using ℓ_1 -norm sparse learning," in *Advances in Neural Information Processing Systems*, pp. 921–928, 2008.
- [31] K. Kowalczyk, E. A. Habets, W. Kellermann, and P. A. Naylor, "Blind system identification using sparse learning for TDOA estimation of room reflections," *IEEE Signal Processing Letters*, vol. 20, no. 7, pp. 653–656, 2013.
- [32] M. Kallinger and A. Mertins, "Multi-channel room impulse response shaping—a study," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 5, pp. V101–V104, 2006.
- [33] A. Mertins, T. Mei, and M. Kallinger, "Room impulse response shortening/reshaping with infinity-and-norm optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 249–259, 2010.
- [34] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multi-channel equalization for speech dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1879–1890, 2013.
- [35] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1818–1829, 2010.
- [36] I. Kodrasi, A. Juki, and S. Doclo, "Robust sparsity-promoting acoustic multi-channel equalization for speech dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 166–170, 2016.
- [37] S. Arberet, P. Vanderghenst, J.-P. Carrillo, R. E. Thiran, and Y. Wiaux, "Sparse reverberant audio source separation via reweighted analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1391–1402, 2013.
- [38] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.
- [39] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.
- [40] X. Li, L. Girin, and R. Horaud, "An EM algorithm for audio source separation based on the convolutive transfer function," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017.
- [41] S. Mirsamadi and J. H. Hansen, "Multichannel speech dereverberation based on convolutive nonnegative tensor factorization for ASR applications," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [42] N. Mohammadiha and S. Doclo, "Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 2, pp. 276–289, 2016.
- [43] D. Baby and H. Van Hamme, "Joint denoising and dereverberation using exemplar-based sparse representations and decaying norm constraint," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 2024–2035, 2017.
- [44] M. Vetterli, "A theory of multirate filter banks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 3, pp. 356–372, 1987.
- [45] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation," *IEEE Transactions on Signal Processing*, vol. 40, no. 8, pp. 1862–1875, 1992.
- [46] J. P. Reilly, M. Wilbur, M. Seibert, and N. Ahmadvand, "The complex subband decomposition and its application to the decimation of large adaptive filtering problems," *IEEE Transactions on Signal Processing*, vol. 50, no. 11, pp. 2730–2743, 2002.
- [47] H. Yamada, H. Wang, and F. Itakura, "Recovering of broadband reverberant speech signal by sub-band MINT method," in *IEEE International*

- Conference on Acoustics, Speech, and Signal Processing*, pp. 969–972, 1991.
- [48] S. Weiss, G. W. Rice, and R. W. Stewart, “Multichannel equalization in subbands,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 203–206, 1999.
 - [49] N. D. Gaubitch and P. A. Naylor, “Equalization of multichannel acoustic systems in oversampled subbands,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1061–1070, 2009.
 - [50] F. Lim and P. A. Naylor, “Robust speech dereverberation using subband multichannel least squares with variable relaxation,” in *European Signal Processing Conference (EUSIPCO)*, 2013.
 - [51] X. Li, L. Girin, and R. Horaud, “Audio source separation based on convolutive transfer function and frequency-domain lasso optimization,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2017.
 - [52] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
 - [53] X. Li, L. Girin, R. Horaud, and S. Gannot, “Estimation of the direct-path relative transfer function for supervised sound-source localization,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 2171–2186, 2016.
 - [54] X. Li, L. Girin, R. Horaud, and S. Gannot, “Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1997–2012, 2017.
 - [55] X. Li, L. Girin, S. Gannot, and R. Horaud, “Non-stationary noise power spectral density estimation based on regional statistics,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 181–185, 2016.
 - [56] D. Campbell, “The roomsim user guide (v3. 3),” 2004.
 - [57] W. G. Gardner and K. D. Martin, “HRTF measurements of a KEMAR dummy-head microphone,” *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.
 - [58] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database,” *National Institute of Standards and Technology (NIST)*, Gaithersburgh, MD, vol. 107, 1988.
 - [59] I. Cohen and B. Berdugo, “Speech enhancement for non-stationary noise environments,” *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.
 - [60] J. F. Santos and T. H. Falk, “Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2197–2206, 2014.
 - [61] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 749–752, 2001.
 - [62] E. Hadad, F. Heese, P. Vary, and S. Gannot, “Multichannel audio database in various acoustic environments,” in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 313–317, 2014.
 - [63] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, *et al.*, “A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research,” *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–19, 2016.